DOCUMENT RESUME

ED 400 292 TM 025 565

AUTHOR Brown, William L.; And Others
TITLE The Reliability and Validity of Mathematics Performance Assessment.
PUB DATE Apr 96
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary School Students; Grade 5; Intermediate Grades; *Interrater Reliability; Item Response Theory; *Mathematics; *Performance Based Assessment; *Problem Solving; Psychometrics; Scoring; Test Reliability; *Test Validity
IDENTIFIERS *FACETS Computer Program; FACETS Model; *Minneapolis Public Schools MN; Rasch Model

ABSTRACT

This study presents psychometric characteristics of the mathematics problem solving performance assessment used in the Minneapolis Public Schools, focusing on the interrater reliability, scoring reliability, and validity of the assessment. The Minneapolis Math Problem Solving Assessment (MPSA) was established in 1991. Students are asked to solve two complex problems, and written communication is a strong component of the assessment. The problems are scored on strategy, quality of solution, math communication, mathematical reflection and connection, and sentence structure, spelling, and mechanics. This study analyzed the 2,600 papers done by about 1,300 students in grade 5 using the FACETS model. FACETS is an extension of the one-parameter item response theory, or Rasch model, which parameterizes each facet of performance assessment into a common scale and enables one to compare elements within one facet and to compare different facets. Using FACETS with classical test theory allowed evaluators to examine interrater reliability, interrater agreement, and validity. Results support the validity and reliability of the MPSA and provide information that can be used to improve the assessment. (Contains 4 figures, 8 tables, and 14 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 400 292

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
William L. Brown
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

The Reliability and Validity of Mathematics Performance Assessment

William L. Brown

Kathryn O'Gorman

Yi Du

Minneapolis Public Schools

BEST COPY AVAILABLE

Paper for Presentation at the Annual Conference American Educational Research Association, New York, 1996

TM025565

Performance Assessment has resurfaced dramatically in the 1990s because it (1) shows promise for assessing learning outcomes that require demonstration of skills or other performances that cannot be assessed using multiple-choice, (2) is linked with teaching and curriculum and (3) relates to real-life skill. But can assessment systems founded on human judgment be demonstrably reliable, valid and fair? Performance assessment challenges psychometricians, test developers and researchers to devise observational situations that evoke evidence about what we want to infer, to learn how to extract and summarize this evidence, and to monitor and improve assessment systems.

This study presents the psychometric characteristics of the math problem solving performance assessment used in Minneapolis Public Schools. Particularly, it focuses on extracting evidence to examine the inter-rater reliability, scoring reliability and validity of this assessment, and on using this evidence to monitor and improve the assessment system.

The Minneapolis Math Problem Solving Assessment

The Minneapolis Math Problem Solving Assessment (MPSA) was established in 1991. It is administered by the MPS Department of Research, Evaluation and Assessment as one of several large-scale performance assessments. The assessment is aligned with the National Council of Teachers of Mathematics (NCTM) Standards, the State of Minnesota Model Learner Outcomes for Mathematics, and the Minneapolis Public Schools' Exit and Supportive Outcomes.

In the assessment, students solve two problems. The problems are complex and require several steps to solve. Students are often asked to find more than one solution, or to find different solutions depending on choices the student makes. Sometimes the student must realize that the solution is dependent on more than one variable. Written communication is a strong component of the assessment.

The problems are scored, on a scale from 1 to 4, on five dimensions:

- ***Problem Solving Strategy(PSS)***: The appropriateness and completeness of strategies
- ***Quality of Solution(QS)***: The accuracy and completeness of the answers
- ***Math Communication(MC)***: The clarity of the solution process and the use of math vocabulary, symbols, and representations (charts, graphs, pictures, etc.)

- ***Mathematical Reflection/Connection(R/C)***: The student's conclusion, generalization, connection to other problems or situations or to a math concept, etc.
- ***Sentence Structure, Spelling, and Mechanics(SS)***: The sentence structure, grammar, spelling, punctuation, etc. of the reflection paragraph

All grade 5 students participated in the MPSA in 1995. A group of Minneapolis teachers was selected to score a representative sample of these papers (about 1300 papers for each of two problems). All raters were trained prior to scoring. Each of the five dimensions was scored on a scale of 1 to 4. Each paper was scored by two raters. If both scores were the same, the paper received that score. Adjacent scores were averaged. If the scores were more than one level apart, a third (master) rater resolved the discrepancy. Approximately 70 percent of the scores were exact matches; about 2 percent required a third rating. Scores were reported for each of the five dimensions.

Data and Methods

This study analyzed the 2600 scored papers done by about 1300 Grade 5 students. Each student responded to two topics: "Job Hunting" and "Tile Trivets." A five-dimensional scoring rubric was used to evaluate the papers.

We analyzed the data using the FACETS model. FACETS is an extension of the one-parameter Item Response Theory (IRT), or Rasch model. FACETS models the probabilities of ordered-category ratings in terms of parameters for students, raters, topics, scoring dimensions and other relevant "facets" of the observation, such as student and rater background variables.

Student parameters capture students' tendencies to receive high or low ratings; rater parameters, their severity or leniency; topic parameters, whether the topics are generally easy or hard; scoring dimension parameters, whether the dimensions are easy or hard. We would expect a lenient rater to give a high rating to a generally high-scoring student on an easy topic according to an easy dimension. Such expectations are formulated more precisely, in probabilistic terms, through the model.

For the math problem solving assessment, the primary FACETS model includes four facets: student, scoring dimension, rater, and prompt:

$$\text{Log}\left(\frac{P_{nijmk}}{P_{nijmk-1}}\right) = B_n - D_i - C_j - A_m - F_{jk} \qquad (1)$$

where P_{nijmk} is the probability of student n being graded in category k by rater j on scoring dimension i and topic m, $P_{nijmk-1}$ is the probability of student n being graded in category k-1 by rater j on scoring dimension i and topic m, B_n is the writing ability measure of student n, D_i is the difficulty calibration of scoring dimension i, C_j is the severity measure of rater j, A_m is the difficulty calibration of prompt k, F_{jk} is the difficulty calibration of grading category k-1 relative to category k. Because each scoring dimension is constructed with its own rating scale, F_{ik} is an item-scale model. The rating scale is k=0,K.

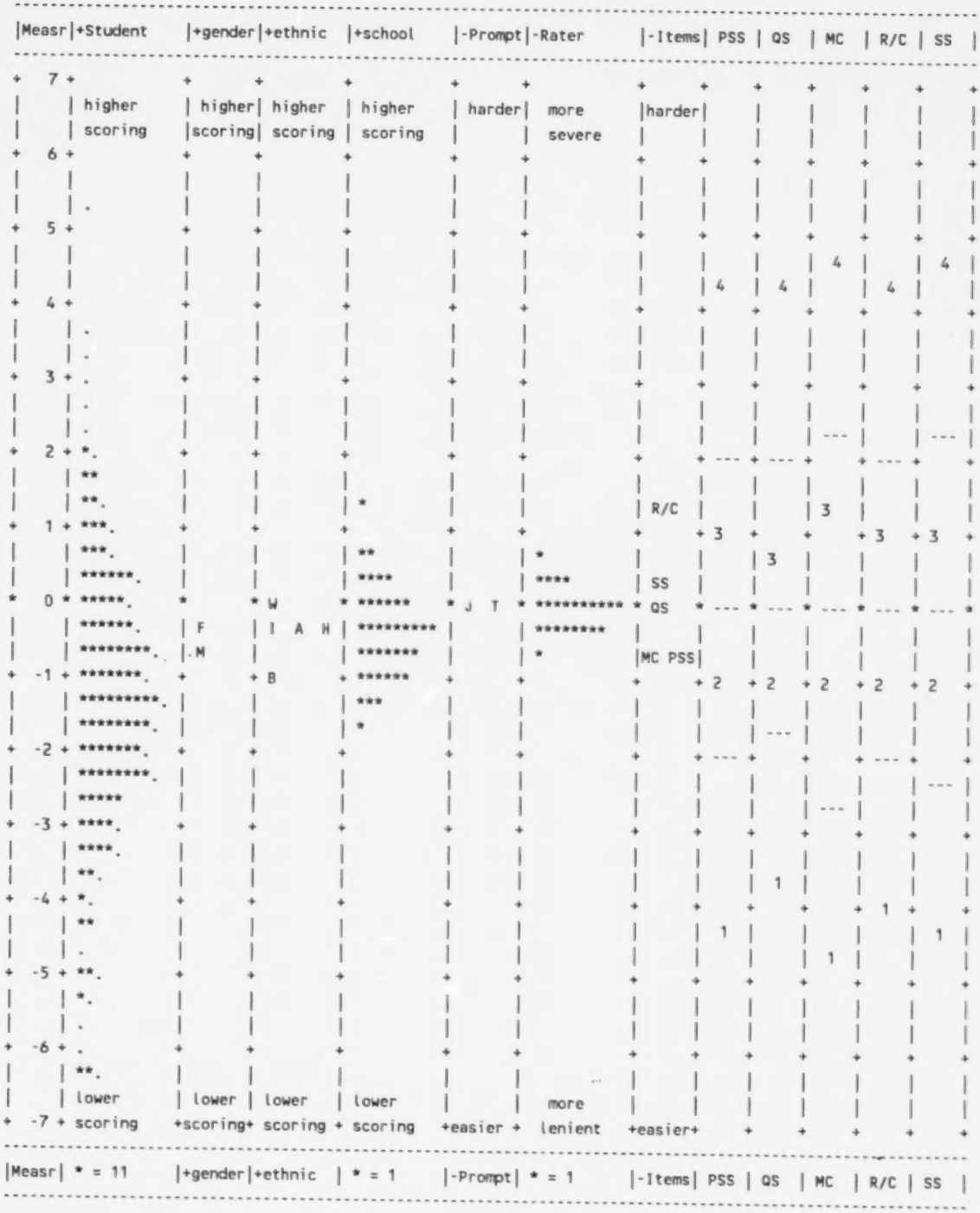
As we will see, typical variation serves as a standard for identifying unusual ratings, papers, raters and so on. With FACETS and Classical Testing Theory, we explored the following questions:

- How accurately are students measured?
- Do the raters differ in the severity with which they rate the MPSA papers?
- Do the raters use the rubric consistently?
- Are the topics difficult for students?
- Are the scoring dimensions valid? Is the difficulty order of the dimensions consistent with the hypothesized expectation?

Results and Interpretation

Figure 1 presents an overview of the estimates of these facets in the assessment. It depicts the estimates for students, raters, topics, scoring dimensions and students' gender and ethnicity along the logit scale. The "raters," "topics," and "scoring dimensions" estimates were specified to center around zero. Note that the students are much more spread out than the raters, with standard deviations of 1.66 and 0.42 respectively.

Figure 1 Map of All Facets



The map shown in Figure 1 enables one to view all facets of the analysis simultaneously, summarizing key information about each facet. It highlights results from more detailed sections of the FACETS output for students, raters, and other facets of the design.

The FACETS program calibrates the raters, students, topics and scoring dimensions so that all facets are positioned on a common scale. That scale is in log-odds, or "logit," units which, under the model, constitute an equal-interval scale with respect to appropriately transformed probabilities of responding in particular categories. Having a single frame of reference for the aspects of the rating process facilitates comparisons within and between facets. The logit scale appears as the **first column** in the map.

The **second column** displays the estimates of students' proficiency on the math problem solving assessment--single-number summaries on the logit scale of each student's tendency to receive low or high ratings, across topics, raters and scoring dimensions. In FACETS terminology, these are "student measures." These student measures are ordered with more proficient students appearing at the top of the column and less proficient students appearing at the bottom of the column. Each star represents 11 students, and a dot represents fewer than 11 students. The student measures range from -7 logits to 7 logits, about a fourteen-logit spread. These measures appear as a fairly symmetrical mesokurtic distribution, looking something like a "bell-shaped" normal curve.

The **third column** compares student gender groups. As this map shows, females performed slightly better than males in the MPSA. The **fourth column** compares student ethnic groups. White students were scored highest, African American students were scored lowest; American Indian, Asian and Hispanic students are in-between. The **fifth column** displays the estimates of schools in the assessment. As with the student estimate, schools with higher scores are on the top, schools with lower scores are at the bottom.

The **sixth column** compares topics used in the assessment in terms of the level of difficulty. The map shows that both topics have similar difficulty levels. A detailed table will be displayed later to show the exact difference between the two topics.

The **seventh column** compares raters in terms of the level of severity or leniency when rating students' papers. In this column, more severe raters appear higher and more lenient raters lower. Rater severity measures ranged from 1 logit to -1 logit. When we compare the distribution of

rater severity measures to the distribution of student measures, the distribution of rater severity measures is much narrower; student measures show a fourteen-logit spread, while rater severity shows a less than a one-logit spread.

The **eighth column** compares scoring dimensions in terms of the level of difficulty. "*Mathematic Reflection/Connection*" (R/C) is hardest. "*Sentence Structure, Spelling, and Mechanics*" is the second hardest. "*Quality of Solution*" is in the middle. "*Mathematical Communication*" and "*Problem Solving Strategy*" are easiest.

The **ninth through thirteenth columns**--one for each scoring dimension--show the most probable rating in each of the scoring dimensions for a student at a given level on the logit scale. The horizontal lines across a column, calculated from the dimension and category parameters, indicate the point at which the likelihood of getting the next higher rating begins to exceed the likelihood of getting the next lower rating. Looking within the PSS column (Problem Solving Strategy), for example, we see that students with measures from -7 logits through -2 logits are more likely to receive a "1" than any other ratings; students with measures between -2 and 0 logits are more likely to receive a "2"; and so on. Looking across columns, we can determine the most likely rating on each of the five dimension scales for a student of a given ability level. For example, a student whose measure was 2.2 logits was most likely to have received a rating of "4" on dimensions 1, 2, and 3 and a rating of "3" on dimensions 3 and 5 (labeled in the map as PSS, QS, MC, R/C, and SS).

Student Measures

Table 1 reports some individual student measures, standard errors, fit statistics and summary statistics for the student facet. The standard errors for student measures average 0.42. The mean of student ability is -1.3, which means that the topics were relatively hard for these students and that raters used relatively strict standards to score students' papers. The reliability of student scores is 0.93. The fit statistics are both 1.0, which means the data fit this model very well. The chi-square statistic, $\chi^2 = 15453.3$ with df = 1161, indicates that these students are significantly different from each other. Another chi-square statistic, $\chi^2 = 1108.1$ with df = 1160, $p < .86$, supports the hypothesis that the distribution of students is normal.

Table 1. Summary of Student Measures

Obsvd Score	Obsvd Count	Obsvd Average	Fair Avrge	Model Measure	Model S.E.	Infit MnSq	Infit Std	Outfit MnSq	Outfit Std	Num	Student
17	10	1.7	1.7	-1.69	0.50	1.2	0	1.0	0	1	10116
28	20	1.4	1.4	-2.48	0.42	0.7	-1	0.6	-1	2	10117
27	20	1.4	1.4	-2.50	0.44	0.9	0	1.1	0	3	12011
19	10	1.9	1.7	-1.58	0.47	0.9	0	0.9	0	4	12212
45	20	2.3	2.2	-0.67	0.32	0.7	-1	0.7	-1	5	12521
31	20	1.6	1.5	-2.09	0.38	1.9	2	2.0	2	6	12679
39	20	2.0	2.0	-1.04	0.33	0.5	-2	0.6	-1	7	12683
23	16	1.4	1.4	-2.36	0.46	1.0	0	0.9	0	8	12703
63	20	3.2	3.2	1.48	0.35	1.4	1	1.3	1	9	12704
61	20	3.0	3.1	1.16	0.34	1.5	1	1.5	1	10	12711
58	20	2.9	3.0	0.99	0.33	0.5	-2	0.5	-2	11	12717
76	20	3.8	3.9	3.84	0.55	1.0	0	1.1	0	12	12718
49	20	2.5	2.5	-0.06	0.32	1.2	0	1.2	0	13	12721
				....							
67	20	3.4	3.5	2.10	0.37	1.0	0	1.1	0	1199	916935
53	20	2.7	2.7	0.39	0.32	0.8	0	0.8	0	1200	917003
28	20	1.4	1.4	-2.63	0.42	1.3	0	1.2	0	1201	917070
20	10	2.0	2.2	-0.61	0.46	0.8	0	0.8	0	1202	92139
53	20	2.7	2.6	0.26	0.32	1.2	0	1.2	0	1203	92310
33	20	1.6	1.7	-1.59	0.36	0.7	-1	0.8	0	1204	92692
54	20	2.7	2.7	0.35	0.33	1.1	0	1.1	0	1205	93503
29	16	1.8	1.7	-1.62	0.38	1.4	0	1.3	0	1206	93538
35	20	1.8	1.7	-1.56	0.35	0.8	0	0.6	-1	1207	93875
59	20	3.0	2.9	0.84	0.34	1.4	1	1.4	1	1208	94673
43	20	2.2	2.1	-0.79	0.32	1.8	2	1.7	2	1209	94679
70	20	3.5	3.6	2.47	0.40	1.0	0	0.9	0	1210	94688
49	20	2.5	2.5	0.02	0.32	3.9	5	3.9	5	1211	94807
49	20	2.5	2.5	0.04	0.32	0.5	-1	0.6	-1	1212	94954
61	20	3.0	3.1	1.17	0.34	1.5	1	1.4	1	1213	94969
45	20	2.3	2.2	-0.62	0.32	0.6	-1	0.6	-1	1214	95083
27	20	1.4	1.3	-2.82	0.44	0.8	0	0.7	0	1215	95101
28	20	1.4	1.4	-2.40	0.42	1.0	0	0.9	0	1216	95121
41	20	2.0	2.0	-1.01	0.33	1.3	1	1.2	0	1217	9697
Obsvd Score	Obsvd Count	Obsvd Average	Fair Avrge	Model Measure	Model S.E.	Infit MnSq	Infit Std	Outfit MnSq	Outfit Std	Num	Student
36.7	18.7	1.9	2.0	-1.20	0.42	1.0	-0.2	1.0	-0.2	Mean (Count: 1215)	
14.1	3.0	0.7	0.7	1.69	0.15	0.5	1.4	0.5	1.4	S.D.	

RMSE 0.44 Adj S.D. 1.63 Separation 3.67 Reliability 0.93
Fixed (all same) chi-square: 15487.1 d.f.: 1161 significance: .00
Random (normal) chi-square: 1108.2 d.f.: 1160 significance: .86

Rater Measures

Figures 2 and 3 plot rater 1's ratings against rater 2's ratings on topics 1 and 2 respectively. The correlation between the two raters on topic 1 is 0.87, and on topic 2 is 0.92. This indicates that the inter-rater agreement between raters 1 and 2 is very high.

Figure 2 First Ratings Vs. Second Ratings on Topic 1 (Tile)

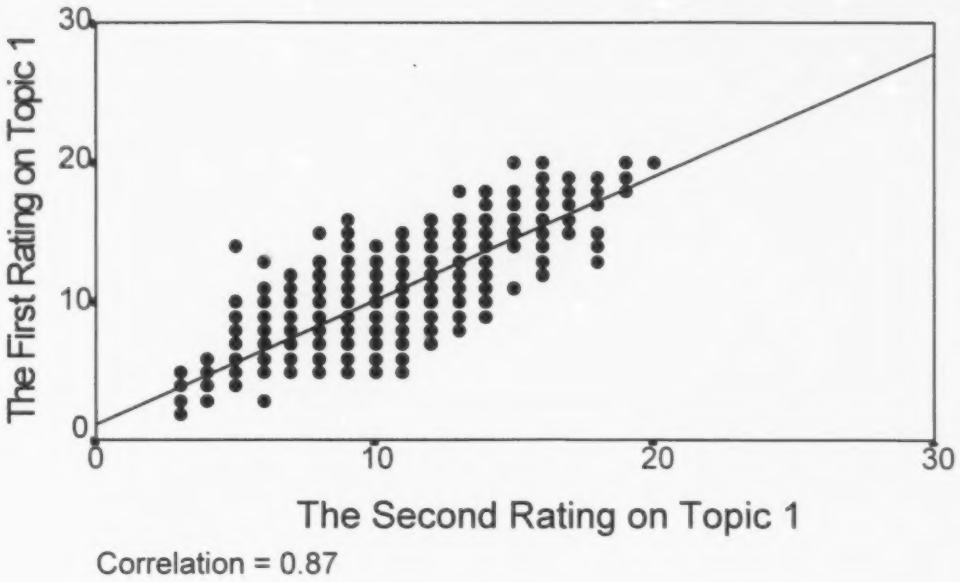
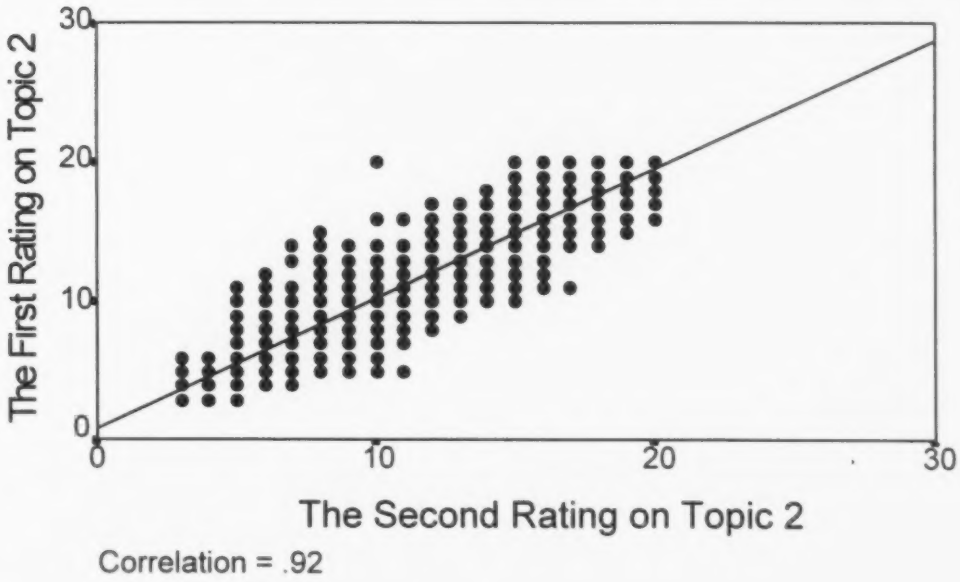


Figure 3 The First Ratings Vs the Second Ratings on Topic 2 (Job)



One important issue in performance assessment is that students should not be disadvantaged if they happened to be rated by severe raters, nor unfairly advantaged if they happened to be rated by lenient raters. Examining discrepant ratings is not an appropriate method for dealing with this problem. Two severe raters may agree in their ratings of a student, but would give significant different ratings than two lenient raters. Finding that raters differ substantially in the degree of severity can suggest a need to address such differences in rater training, or to consider the feasibility of adjusting students' scores in accordance with the severity or leniency of the ratings.

FACETS produces a measure of the degree of severity each rater, accounting for the various mixtures of high-rated and low-rated papers. Table 2 orders the rater severity measures from the most lenient at the top to the most severe at the bottom, in the column labeled "rater measures." To the right of each Rater Measure is the standard error of the estimate, indicating the precision with which it has been estimated. The rater severity measures ranged from -0.51 to 0.78, a 1.29 logits spread.

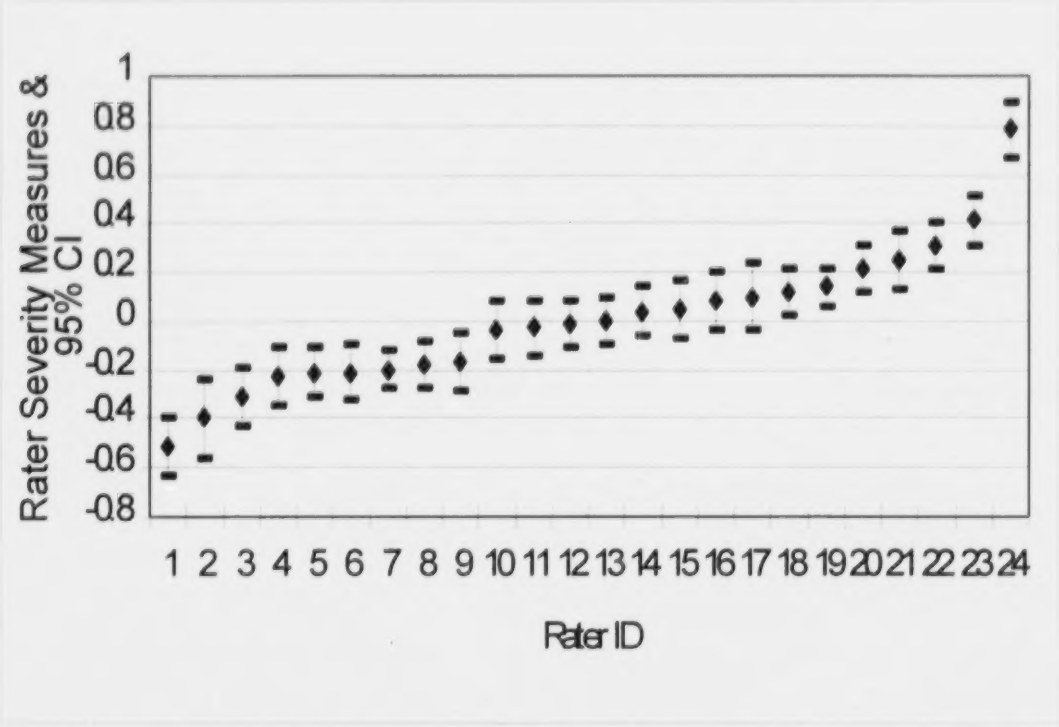
A chi-square test of the hypothesis that the raters all have the same degree of severity in ratings has a value of 569 on 23 degrees of freedom, meaning there is virtually no chance that the differences among estimated rater severity measures would have arisen from identically severe raters. It is not surprising that this difference is statistically significant, because so many papers were rated (about 2600). Figure 4 plots rater severities and its 95% confidence intervals which show that most raters do not have significant differences in terms of severity.

Table 2 Rater Severity Measures

	Rater ID	Number of Rating	Rater Severity Measure	S.E.	Infit Mean Square	Outfit Mean Square
Most	1	813	-0.51	0.06	1.0	1.2
Lenient	2	458	-0.40	0.08	1.0	1.0
	3	785	-0.31	0.06	0.9	0.9
	4	807	-0.23	0.06	1.1	1.2
	5	952	-0.21	0.05	1.0	1.0
	6	800	-0.21	0.06	1.1	1.0
	7	1256	-0.20	0.04	1.0	1.0
	8	900	-0.18	0.05	0.9	0.9
	9	743	-0.17	0.06	1.0	1.1
	10	646	-0.04	0.06	0.9	0.9
	11	621	-0.03	0.06	0.8	0.9
	12	859	-0.01	0.05	0.9	0.9
	13	997	0.00	0.05	0.9	0.9
	14	1113	0.04	0.05	1.1	1.0
	15	898	0.05	0.06	0.9	0.9
	16	784	0.08	0.06	1.3	1.4
	17	547	0.10	0.07	1.1	1.1
	18	948	0.12	0.05	1.1	1.0
	19	1692	0.14	0.04	0.9	0.9
	20	1166	0.21	0.05	1.0	1.0
	21	927	0.25	0.06	1.0	0.9
	22	954	0.31	0.05	1.1	1.0
Most	23	1272	0.41	0.05	1.0	0.9
Severe	24	937	0.78	0.06	1.0	0.8
	Mean	911.5	0.0	0.1	1.0	1.0
	S.D.	182.3	0.2	0.0	0.1	0.1

Figure 4 shows rater severity levels and their 95% confidence intervals. When the confidence intervals do not overlap, a significant difference between rater severity exists. We can identify very lenient and very severe raters, using these results, for further training and rechecking student scores.

Figure 4. 95% Confidence Intervals around Rater Severity



As an illustration, Table 3 gives probabilities for ratings on Dimension *"Problem Solving Strategy (PSS)"* from an average student as rated by each of the raters, from the most lenient to the most severe. The most likely rating for most raters is a "2", with a probability ranging from 0.4 to 0.6 in most cases. The most severe raters will rate a "2" paper on this dimension a "1."

Table 3 Probabilities of Raters' Rating for Scoring Dimension

Problem Solving Strategy (PSS)

	Rater ID	Rater Severity Measure	Probability of Rating Category				Expectation	Most Likely Rating
			1	2	3	4		
Most	1	-0.51	0.13	0.57	0.25	0.05	2.22	2
Lenient	2	-0.40	0.18	0.55	0.22	0.05	2.13	2
	3	-0.31	0.23	0.53	0.20	0.04	2.04	2
	4	-0.23	0.28	0.51	0.17	0.04	1.96	2
	5	-0.21	0.30	0.51	0.16	0.03	1.93	2
	6	-0.21	0.30	0.51	0.16	0.03	1.91	2
	7	-0.20	0.31	0.51	0.16	0.02	1.89	2
	8	-0.18	0.33	0.50	0.15	0.02	1.86	2
	9	-0.17	0.34	0.50	0.15	0.01	1.83	2
	10	-0.04	0.42	0.47	0.11	0.01	1.70	2
	11	-0.03	0.43	0.47	0.11	0.00	1.68	2
	12	-0.01	0.45	0.46	0.10	0.00	1.64	2
	13	0.00	0.45	0.46	0.10	0.00	1.65	2
	14	0.04	0.47	0.45	0.08	0.00	1.61	2
	15	0.05	0.48	0.44	0.08	0.00	1.60	2
	16	0.08	0.50	0.43	0.07	0.00	1.57	2
	17	0.10	0.51	0.43	0.06	0.00	1.55	2
	18	0.12	0.52	0.42	0.06	0.00	1.53	2
	19	0.14	0.54	0.42	0.05	0.00	1.51	2
	20	0.21	0.58	0.39	0.02	0.00	1.44	1
	21	0.25	0.61	0.38	0.01	0.00	1.40	1
	22	0.31	0.64	0.36	0.00	0.00	1.36	1
Most	23	0.41	0.68	0.32	0.00	0.00	1.32	1
Severe	24	0.78	0.83	0.17	0.00	0.00	1.17	1

FACETS produces measures of within-rater consistency for each rater in terms of fit statistics: infit and outfit. The infit statistic is unstandardized, which detects inconsistent rating behavior using the rating criteria; the outfit is a standardized outlier-sensitive fit statistic that is strongly influenced by single unexpectedly large residuals. The mean square fit statistics are with expectation 1 and range 0 to infinity.

The infit and outfit mean square values less than 1 indicate less variation than average in a rater's ratings, which means "too much consistency." Often the problem is that the rater fails to use all the points on the scale. The infit and outfit mean square greater than 1 indicates greater than average variation in a rater's ratings. This rater is not using the rating scales in as consistent a manner as the others. The fit values between .7 to 1.5 (Wright and Linacre) indicate internally consistent and appropriate ratings. The rater fit mean-square statistics shown in Table 2 range from 0.8 to 1.4, suggesting that all 24 raters were internally consistent in their ratings.

Scoring Dimension Measures

Table 4 shows the scoring dimension calibrations, the standard errors of these estimates, and the mean-square fit statistics for each scoring dimension. *Mathematical Reflection/Connection* was expected to be the most difficult dimension. Students have the least experience with it and it is the most abstract part of the assessment. *Sentence Structure, Spelling and Mechanics* was expected to be easier than *Reflection/Connection* but still somewhat difficult because it is connected to the reflective task. The Quality of Solution was expected to be more difficult than *Problem Solving Strategy*, since the solution flows from the Strategy but also takes into account computation and other errors in the execution of the strategy. The calibrated difficulty order of the results are consistent with these hypothesized expectation. The dimensions fit statistics verify that these scoring dimensions, in general, fit the measurement model.

Table 4 Scoring Dimension Calibration

	Dimension	Score	Count	Calibration	SE	MnSq	MnSq
Hard	R/C	6522	4336	1.38	0.03	1.1	1.1
	SS	8360	4336	0.07	0.02	1.1	1.2
	QS	9089	4678	-0.05	0.02	1.1	1.0
	MC	10060	4677	-0.59	0.02	0.9	0.8
	PSS	10454	4678	-0.80	0.02	0.9	0.9
	Mean	8720	4382	0	0.02	1	1
	SD	1373	143	0.76	0.00	0.1	0.1

Tables 5 and 6 report the correlation coefficients between the scoring dimensions on the two topics, respectively. The results show that the first three scoring dimensions (*Problem Solving Strategy, Quality of Solution, Math Communication*) are highly correlated and appear to measure one thing. However, the dimensions (*Mathematical Reflection/Connection* and *Sentence Structure, Spelling*, and *Mechanics*) do not correlate well with the first three dimensions, although the other two dimensions are correlated with each other well.

We wouldn't expect *Sentence Structure*, *Spelling* and *Mechanics* to correlate highly with *Problem Solving Strategy, Quality Solution* and *Mathematical Communication*, since it measures traditional writing skills, not mathematical skills. It was included in the scoring because the developers wanted to emphasize the relevance of "writing across the curriculum."

Table 5 Correlation Coefficients of Scoring Dimensions on Topic "Job Hunting"

	PSS	QS	MC	R/C	SS
PSS	1.00	0.91	0.82	0.49	0.49
QS		1.00	0.78	0.48	0.47
MC			1.00	0.52	0.55
R/C				1.00	0.77
SS					1.00

Table 6 Correlation Coefficients of Scoring Dimensions on Topic "Tile Trivets"

	PSS	QS	MC	R/C	SS
PSS	1.00	0.89	0.84	0.55	0.52
QS		1.00	0.77	0.53	0.48
MC			1.00	0.61	0.60
R/C				1.00	0.80
SS					1.00

Prompt Measures

Table 7 shows the prompt calibrations, the standard errors of these estimates, and the mean-square fit statistics for each prompt. Prompt *Job Hunting* is more difficult than *prompt Tile Trivets*. The fit statistics for both prompts are between 1.0 and 0.9 which indicates that the data from the prompts fit the model well enough to function as useful prompts for measuring student math problem solving proficiency. The difficulty difference between the prompts is significant, $\chi^2(1) = 143.6$, $p < 0.001$ with a high separation reliability (0.99). This indicates that "Job Hunting" is significantly more difficult than "Tile Trivets."

Table 7 Prompt Measurement

Obsvd Score	Obsvd Count	Obsvd Average	Fair Avrge	Model Measure	Model S.E.	Infit MnSq	Infit Std	Outfit MnSq	Outfit Std	N	Prompt
22002	10776	2.0	2.6	-0.13	0.02	1.1	4	1.0	2	2	Tile trivets
21603	11139	1.9	2.4	0.13	0.02	0.9	-5	0.9	-3	1	Job hunting
21802.5	10957.5	2.0	2.5	0.00	0.02	1.0	-0.4	1.0	-0.4	Mean (Count: 2)	
199.5	181.5	0.1	0.1	0.13	0.00	0.1	4.9	0.1	2.9	S.D.	

RMSE 0.02 Adj S.D. 0.13 Separation 8.41 Reliability 0.99
Fixed (all same) chi-square: 143.6 d.f.: 1 significance: .00

Rating Scale

The rating scale score categories and the frequency with which each score was assigned are shown in Table 8 for scoring dimensions. The categories 1 and 2 were used much more heavily than categories 3 and 4. We interpret this difference to be the result of two factors: (1) this scale was designed to be used for both Grades 5 and 7, and (2) these skills are not yet fully taught. As

a developmental scale, it seems reasonable that most Grade 5 students were clustered into categories 1 and 2.

Table 8 Rating Scale

Scoring Category	Problem Solving Strategy Percentage	Quality of Solution Percentage	Math Communication Percentage	Reflecting on the Problem Percentage	Sentence Structure, Spelling and Mechanics Percentage
1	30%	45%	26%	62%	36%
2	29%	24%	41%	28%	40%
3	29%	22%	25%	8%	20%
4	12%	9%	8%	2%	4%

Conclusion

The results of this study support the validity and reliability of the MPS Math Problem Solving Assessment. This study also demonstrates the advantages of the combination of the FACETS model and Classical Testing Theory to construct reliable and valid performance assessment. FACETS parameterizes each facet of performance assessment (students, raters, items, prompts, and so on) into a common scale and enables one to compare elements within one facet and to compare different facets. By using FACETS along with the Classical Test Theory, we are able to examine the interrater agreement, interrater reliability and validity of the math problem solving assessment.

Generally, any performance assessment requires examination of raters, examinees, prompts and scales. Any performance assessment requires examination of reliability and validity. With multiple psychometric methods, it is possible to extract evidence to examine the inter-rater reliability, scoring reliability and validity of this assessment, and to use this evidence to monitor and improve the assessment system.

REFERENCES

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. Journal of Educational Statistics, 13, 1-18.

Breland, H. M., & Gaynor, J. L. (1979). A comparison of direct and indirect assessment of writing skill. Journal of Education Measurement, 16, 119-128.

Engelhard, J. G. (1992a). The measurement of writing ability with a many-faceted Rasch model. Applied Measurement in Education, 5, 171-191.

Linacre, J. M. (1987). The calibration of essay graders. Paper presented at the Midwest Objective Measurement Seminar, Chicago.

Linacre, J. M. (1988). FACETS, Chicago: MESA Press.

Linacre, J. M. (1989). Many-facet Rasch Measurement. Chicago, IL: MESA Press.

Linacre, J. M. (1994a). Constructing measurement with a many-facet Rasch model. In Wilson M. (Ed.), Objective Measurement, Theory into Practice, 2, (pp. 129-144). Norwood, NJ : Ablex Publishing Co.

Linacre, J. M. (1994b). Measurement of judgment. International Encyclopedia of Education (2nd. ed.) Oxford: Pergamon Press.

Linn, R. L. & Burton, E. (1994). Performance-based assessment: implications of task specificity. Educational Measurement: Issues and Practice, 13, 5-8.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley Publishing Company.

Myford, C. M. & Mislevy, R. J. (1995). Monitoring and Improving a Portfolio Assessment System. Princeton: Educational Testing Service.

Wright, B. D. (1991b). Scores, reliabilities and assumptions. Rasch Measurement SIG Newsletter, 5(3), 157-158.

Wright, B. D., & Linacre M. (1994). Reasonable Mean-square Fit Values. Rasch Measurement SIG Newsletter, 8(3), p. 370.

Yen, W. M. (1993). The Maryland school performance assessment program: Performance assessment with psychometric quality suitable for high stakes usage. Paper presented at the Large Scale Assessment Conference. Albuquerque, NM.